DOCUMENT RESUME

ED 322 201 TM 015 404

AUTHOR Nolet, Victor; Tindal, Gerald
TITLE Evidence of Construct Validity in Published Achievement Tests.
PUB DATE Apr 90
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *Construct Validity; Elementary Secondary Education; *Norm Referenced Tests; Scoring; Standardized Tests; *Test Interpretation; Test Results; Test Use; Test Validity
IDENTIFIERS *Test Batteries; Test Publishers

ABSTRACT

Valid interpretation of test scores is the shared responsibility of the test designer and the test user. Test publishers must provide evidence of the validity of the decisions their tests are intended to support, while test users are responsible for analyzing this evidence and subsequently using the test in the manner indicated by the publisher. Publishers of achievement batteries provide a variety of types of data to support the technical adequacy of their tests; however, the utility of this information as evidence of construct validity has not been explored. This study analyzed data provided by achievement test publishers to investigate the existence of a network of evidence to support various inferences about the meaning of scores obtained from these measures. Focus was on examining various aspects of the construct validity of published norm-referenced achievement test batteries. Nine group and four individually administered achievement tests were reviewed. The materials examined for each test included student response booklets, scoring protocols, administration manuals, objectives lists, test coordinators' handbooks, and technical manuals. Results show that achievement test batteries are adequate measures of general achievement in the broadly defined constructs of reading, mathematics, and language expression; however, inferences about student performance in skill areas represented by the various subtests included in most achievement batteries seem not to be supported. It is concluded that test publishers are ill-advised to demarcate many subskills in the categories of reading, language, and mathematics. Reliability and validity data indicate that the fewer the number of facets into which constructs are divided, the better. The published achievement test batteries studied seem to have convergent validity but no discriminant validity and mono-operation bias, severely limiting the kinds of inferences that can be made. These tests fail to represent the wide range of classroom-relevant behaviors that are components of each construct. Four tables are included. (RLC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

VICTOR NOLET

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED322201

Evidence of Construct Validity in Published Achievement Tests

Victor Nolet

Gerald Tindal

Running Head: Validity in Published Achievement Tests

Paper presented at the American Educational Research Association,

Boston, MA, April 16-20, 1990

Requests for reprints: Dr. Gerald Tindal, 275 Teacher Education,

University of Oregon, Eugene, OR. 97405-1213

BEST COPY AVAILABLE

TM015404

Abstract

Valid interpretation of test scores is the shared responsibility of the test designer and the test user. Test publishers must provide evidence of the validity of the kinds of decisions their tests are intended to support while test users are responsible for analysis of this evidence and subsequent use of the test in the application indicated by the publisher. Publishers of achievement batteries provide a variety of types of data to support the technical adequacy of their tests, however, the utility of this information as evidence of construct validity has not been explored. This study involved analysis of data provided by achievement test publishers to investigate the existence of a network of evidence to support various inferences about the meaning of scores obtained from these measures. Achievement test batteries are characterized as adequate measures of general achievement in the broadly defined constructs of reading, mathematics and language expression, however, inferences about student performance in skill areas represented by the various subtests included in most achievement batteries seem not to be supported.

Introduction

The technical adequacy of published norm-referenced achievement tests to make educational decisions has been questioned frequently (Ebel, 1978; Salvia & Ysseldyke, 1988); as a consequence an extensive literature on their validity has emerged. This research has focused on the manner and extent to which published achievement tests sample a specified domain of interest, a characteristic of tests generally labeled content validity.

Published achievement tests consistently have been shown to lack content validity and perform poorly as curriculum-referenced measures of achievement . For example, one finding that has been reported and replicated is that inferences about student performance in a curriculum are dependent upon the specific achievement test used as a dependent measure (Jenkins & Pany 1978; Shapiro & Derr, 1987). Similarly, it has been shown that significant differences in test performance can be predicted by differences in test-curriculum overlap (Good and Salvia, 1988) and although the amount of overlap between tests and curricula often is minimal, this overlap is a powerful predictor of end-of-year test performance (Leinhardt, 1983). Finally, although published curricula include most of the topics found on published achievement tests, the tests tend not to be representative samples of the topics presented in curricula (Freeman, Kuhs, Porter, Floden, Schmidt, & Schwille, 1983).

Validity involves an evaluation of the extent to which multiple lines of evidence support inferences about the scores that result from an assessment procedure (Messick, 1989). Inasmuch as content validity refers to the relevance and representativeness of a test with respect to a particular domain, it has more to do with test construction than with inferences about scores obtained from the test and may not be a form of validity at all (Messick, 1981; 1989). The finding that published achievement tests lack content validity should not be surprising because these tests are not constructed sample a domain of any curriculum. In fact, many test publishers are explicit in their creation of tests that sample from a very broadly defined domain such as "the basic curricular content taught nationwide" (Prescott, Balow, Hogan & Farr, 1988, p 9) rather than a particular curriculum program. In effect, the research findings of poor content validity for published achievement tests may have merely provided empirical validation of the test

construction procedures described in most test technical manuals. Unfortunately, with so much attention paid to content validity, the more fundamental issue of the construct validity of published achievement tests largely has been ignored.

According to Cronbach & Meehl (1955), "Construct validity is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined" (p 282). Construct validation refers to development of a body of evidence that supports specific inferences about the meaning of scores obtained from a particular test. Such inferences are more or less valid, depending on the meaning ascribed to the construct the test is intended to measure. In this respect, validity refers to the inferences that can be made on the basis of a test, not the test itself. Construct validity, then is the shared responsibility of the test designer and the test user. The test designer is responsible for defining the construct and providing evidence that the test adequately measures it and the test user must decide how to interpret scores obtained from the test on the basis of evidence provided by the designer.

Evidence to support construct validity falls into two categories: convergent validity and discriminant validity. Convergent validity involves multiple forms of evidence that show a test adequately measures the construct of interest. As noted by Campbell & Fiske ,(1959) "Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods" (p 83). The discriminant validity of a test would be supported by evidence that variables related to constructs other than the one the test is intended to measure do not influence scores on the test. In this respect, construct validation is a process of showing what a test measures as well as what it does <u>not</u> measure.

Ideally, the shared responsibility for validating a test would suggest that users and developers enter into an implicit bargain. The test publisher "agrees" to provide information that is sufficiently complete and appropriate for the user to make an informed decision about the kinds of interpretations that can be made from the test and the test user "agrees" to be sufficiently knowledgeable and disposed to evaluate the information provided.

The extent to which test users have met their end of the bargain has been well documented. The literature regarding test use has provided a clear indication that teachers make little use of

published tests and probably lack even rudimentary knowledge necessary to evaluate the kind of technical information that may be associated with interpreting test scores (Lazar-Morrison, Polin, Moy & Burry, 1980). One might speculate that teachers' failure to employ published achievment test scores for making instructional decisions (Salmon-Cox, 1981) because these tests, relying almost exclusively on selection type responses, sample behaviors that are distal to daily classroom activities. The overwhelming evidence of the poor content validity of these measures would certainly support this speculation. Similarly, lack of sufficient test and measurement knowledge among test users might imply that not only are the scores obtained from these tests of little value to teachers, the technical data provided by publishers to support inferences about these scores may be of minimal utility.

If test users are unprepared to meet their responsibility to evaluate the validity of their inferences about test scores, to what extent have test publishers have fulfilled their end of the bargain to provide users with adequate information?. Clearly, research indicates the content of these tests may be irrelevant and therefore, of limited use for making inferences relative to classroom instruction. Similarly, although most test publishers provide some indication of criterion related validity, evidence that scores from one test battery correlate highly with scores from another similar battery may not be sufficient for demonstrating the existence of a network of relationships necessary to support specific inferences about the meaning of test scores in applied situations. As Messick (1981) has noted, the various "types" of validity are not comparable and inferences about the meaning of test scores depends on evidence for construct validity, which might include but is not limited to criterion correlations.

To date, little systematic investigation of the construct validity of published achievement tests has been conducted. Previous studies have indicated that test publishers often fail to report sufficient data to support the construct validity of their tests (Hall, 1985; Petrosko, 1978) however, the quality of data provided by publishers has not been addressed. The absence of empirical validation of the construct validity of published achievement tests may result in uncritical acceptance of the assertions of test publishers that a range of inferences regarding the meaning of scores obtained from their tests are supported. The purpose ot the current study is to examine

the extent to which test publishers provide evidence of the contstruct validity of their tests. To examine construct validity, the structure of the tests must be examined.

The Structure of Published Achievement Batteries

The assumptions that apparently underlie construction of published achievement tests have direct bearing on the kind of evidence required to support their convergent and discriminant validity. The tests of interest here involve multiple subtests intended to measure a variety of broad achievement constructs such as reading, math or written expression. The premises underlying the format and construction of these batteries seems to be that (a) subtests with different names test different skills, (b) different subtests may measure separate facets of a single construct, and (c) specific batteries may vary with respect to the number and names of subtests they contain but, presumably, all batteries sample from a similar universe of academic skills or constructs and the name of a subtest has meaning with respect to that universe.

A fundamental assumption underlying the inclusion of multiple tests in achievement batteries is that tests with different names measure distinctly different constructs or facets of a single construct. For example, tests of reading achievement are considered tests of the construct "reading" while mathematics subtests are thought to sample the distinctly different construct "mathematics achievement". Similarly, the validity of inferences about scores obtained from tests named "reading comprehension" and "listening comprehension" will be diminished or enhanced by the extent to which the two subtests sample distinctly different behaviors.

A further assumption in the construction of achievement batteries is that few constructs of interest in school learning are unitary. For example, the construct "reading", often is characterized as involving at least three facets; reading comprehension, reading fluency and vocabulary (Farr & Carey, 1986). Similarly, written language might involve variables such as: syntax, semantics and spelling (Issaacson, 1985). In most batteries, multiple tests are used to measure achievement in various facets of multi-dimensional constructs. For example. such batteries typically include tests of reading comprehension and word attack to test the construct "reading"; math computation and math problem solving to measure the construct "mathematics achievement", and language usage and language mechanics to measure the construct "written expression".

Finally, the validity of inferences about scores obtained from a test relies on the extent to which the test measures the trait of interest rather than error, i.e., the test's reliability. Campbell and Fiske (1959) suggest that reliability and validity represent different points on a single continuum in that both require agreement between two or more measures. The type of reliability required will depend directly on the kind of interpretations the test is intended to support. For example, evidence of test-retest or alternate forms reliability must be provided for tests intended to be used as measures of growth over time in either pre-post or parallel forms administrations. Minimally test publishers would be expected to provide adequate evidence of the the internal consistency of tests.

The current study was undertaken to investigate and describe number of variables associated with the quality of information provided by test publishers as evidence of the construct validity of their tests. The questions investigated were:

1. Do publishers provide sufficient evidence of the reliability of their tests?
2. What skills or constructs do achievement test batteries sample?
3. Do subtests sample the constructs they are intended to measure?
4. Do the patterns of intercorrelations among subtest scores on batteries support the convergent and discriminant validity of the tests?

Method

This study was undertaken to examine various aspects of the construct validity of published norm-referenced achievement test batteries. Underlying the procedures employed in this investigation was the assumption that validation of a test is the shared responsibility of a test publisher and a test user. Publishers bear responsibility for supplying evidence to support the technical adequacy of their instruments while test users are required to evaluate data supplied by the publisher to decide whether a particular test application is valid and ethical. Therefore, only data sources made available by test publishers, for use by test consumers, were examined in this study.

Materials

Thirteen published norm-referenced achievement tests widely used in testing programs in educational settings were analyzed. Nine group and four individually administered achievement tests were reviewed (see Appendix A). With one exception, tests that had national normative data no more than 10 years old at the time of the study were used. The latest version of the Woodcock-Johnson Psych-Educational Test Battery was not available when the study was conducted and the version examined is based on 12 year old norms.

The materials examined for each test included student response booklets, scoring protocols, administration manuals, objectives lists, test coordinators handbooks, and technical manuals. Most of the materials examined were included in specimen kits obtained from test publishers. However, technical manuals containing reliability and validity information typically were not included in these kits and had to be ordered separately. All levels of each test were examined. When multiple forms of a test were available, the version that included levels across the most age or grade ranges was selected.

This study was concerned with achievement test batteries intended to assess multiple constructs so tests aimed at single skills such as oral reading, written language or mathematics were not reviewed. Similarly, within batteries, only tests and subtests focusing on the basic skill areas of listening comprehension, reading, language arts, and mathematics were analyzed. Content area subtests (such as science or social studies) were not evaluated to maintain a focus on basic skills and due to the inconsistent inclusion of these subtests within many batteries. Study or reference skills subtests were evaluated because they were judged to assess skills primarily related to reading rather than a particular curriculum content area.

Procedures

Analyses focused on three sources of evidence of the validity of each test. First a general index of reliability was obtained for each battery on the premise that unreliable tests cannot be considered valid measures (Tindal & Marston, 1990). Second to assess the extent to which test names represent the behaviors they actually sample, individual test items were sorted into nine skill categories, standardized across all batteries. Finally, intercorrelation among subtests within

each battery were examined to verify the existence of a pattern in which subtests intended to measure facets of the same construct are more highly correlated with one another than they are with subtests intended to measure distinctly different constructs.

Reliability Estimates

Generally, tests with reliability estimates below .80 have been considered inadequate for educational decision making, including the kinds of screening decisions for which published, norm-referenced achievement tests are appropriate (Nunnally, 1967; Salvia & Yssledyke, 1988; Webb, 1983). Therefore, .80 was selected as a cutoff score for the purposes of the current study.

Test batteries reported a number of types of reliability, including test-retest, internal consistency, item response, and alternate forms, with considerable variability in the combination of types reported by any one test. Internal consistency reliability estimates (using the KR-20 formula) were reported by most of the batteries examined and, therefore, this type of reliability was chosen for analysis.

In each battery, all KR-20 reliability coefficients reported for all levels of each subtest were examined and the percent of coefficients below .80 was counted. When more than one norming was reported (for example, spring and fall), coefficients from both normings were included. Only subtest reliabilities were observed, so no cluster or composite test scores such as Total Test or Total Battery scores were included in this analysis.

Item Classification

To analyze of skill domains, the total number of items in each test battery were counted within each of nine areas. To standardize the comparison, the nine skill domains were defined as follows:

Reading Comprehension — Knowledge of word meanings, vocabulary knowledge or word, sentence or passage comprehension.

Reading Decoding	Print translation, including phonics, syllabication, structural analysis, and construction of compound words.
Study Skills	Dictionary or index skills, alphabetizing or skills in obtaining information from maps, charts, graphs and tables.
Language Mechanics	Knowledge or use of rules related to grammar, punctuation, capitalization, verb tense, noun-verb agreement or parts of speech.
Language Expression	Knowledge of qualitative characteristics of writing, including style, clarity, or composition were classified in this category. These included items that test sentence order in paragraphs, word order in sentences, main idea, sentence fragments, and run-on sentences.
Spelling	Knowledge of correct letter order in words or items that test knowledge of spelling conventions and rules.
Listening Comprehension	Items presented orally by the teacher, that require students to draw conclusions, make inferences or predictions were included in this category. Response demands could include selection of picture, letter or words.
Mathematics Applications	Knowledge of measurement, math or geometry vocabulary, solving word problems or interpreting charts or graphs through use of mathematical operations, knowledge of number names, counting, number order, place value, expanded notation and number theory.
Mathematics Computation	Solution of problems in which only the number problem is provided. Items may involve math or geometry operations, estimation, math facts, proportions or working with fractions (for example Least Common Denominator or Greatest Common Multiple).

Most of the tests batteries included a list of skill objectives and the numbers of items that tested those objectives. These objectives were examined and classified into one of the nine categories and all items testing the objectives were placed in the corresponding category. When skill objectives were unavailable, the actual test items were examined and classified.

The reliability of the objectives classification procedure was established as follows. Trained data collectors categorized objectives sampled randomly from different test batteries. They then classified specific items associated with each objective (as specified by the test publisher). When classifying specific items, the data collectors were blind to the objective with which each item was associated. The percent of specific items that were coded in the same category as the objectives they were intended to measure was 96%.

All data collectors (3 graduate students in a graduate Teacher Education program at a medium size univerisity) received at least 2 hours of training and then item classifications were conducted. Inter-rater agreement for classification of objectives and specific items was above .90 for all nine skill categories. Observer drift was monitored with periodic reliability checks throughout the classification process and discussion of disagreements in classifications. Then two types of data were obtained. First, the percent of items in each battery placed in each category was computed Second, an index of the congruence of the behaviors sampled by subtests and the skills they purported to measure was developed.

To accomplish this second analysis, subtests were placed in one of the four clusters: listening comprehension, reading, language or math. Generally, subtests that were specified as facets of a particular construct by the test publisher were considered grouped appropriately. For example, if in a particular battery subtests labeled Vocabulary, Word Recognition, and Reading Comprehension contributed to a Total Reading score, these subtests were placed in the Reading subtest cluster. With the exception of subtests that measured study skills, most subtests were clearly identified by publishers as belonging to one of the four clusters. When study skills subtests were not grouped by a publisher in a particular cluster, they, were evaluated as a component of reading since they are primarily related to the skill of reading (often including

analyses of reading and reference materials), rather than a particular curriculum content area (Farr & Carey, 1986).

The nine skill classifications were grouped into construct categories as follows: The construct Listening Comprehension consisted of the Listening Comprehension (LC) skill category only.The construct Reading was comprised of the Reading Decoding (RD), Reading Comprehension (RC) and Study Skills (SS) categories. The construct Language was comprised of the Language Mechanics (LM), Language Expression (LE) and Spelling (SP) skills categories. Finally, the construct Mathematics consisted of the Math Computation (MC) and Math Applications (MA) categories. The percent of items from each subtest cluster placed in each construct category was determined and the percent of items placed in an appropriate construct categories was computed. For example when an item from a subtest in the Reading subtest cluster (for example "Word Recognition") was placed in one of the skill categories that comprised the Reading construct (RD, RC or SS), it was considered an appropriate categorization.

Patterns of Intercorrelations

The procedures used in this analysis have been described elsewhere (Nolet &Tindal, 1990) and will only be briefly summarized here. On each battery, subtests were sorted into one of three categories: reading-related, language-arts-related, and mathematics-related. For example if the subtests labeled Mathematics Computation and Mathematics Application contributed to a Total Math score in a particular battery, these subtests were considered related. All subtests that were not identified by the publisher as a measure of a particular construct were considered unrelated. For example, any subtest that was neither Mathematics Computation nor Mathematics Application would be considered unrelated to either of these math subtests.

Finally, the intercorrelation of all subtests within each battery were examined. All correlation coefficients reported for each subtest across all levels of the battery were analyzed and for each subtest two distributions of correlation coefficients were developed; those associated with related subtests and those associated with unrelated subtests. Therefore six distributions were created for each battery, (i.e., related and unrelated for reading, language arts, and math). The range and median were computed for each of these distributions. The median *related* coefficient was

compared with the median *unrelated* coefficient and the amount of overlap between the two distributions was assessed. Visual comparison of medians and ranges was accomplished through use of Tukey's notched box plots (1977). This method allowed estimation of the significance of difference between medians as well as comparison of overlap of related and unrelated distributions.

Results

Results of the various analyses are shown in Tables 1 through 4 with group administered tests displayed in the top portion and individually administered tests placed in the lower portion of each table.

Reliability Estimates

Test Battery names are shown in the left column of Table 1, the total number of reliability coefficients reported (KR-20) for each battery is shown in the middle column, and the percent of these below .80 is shown in the right column. . Across batteries, the percent of coefficients below .80 ranged from 0 on the TAP and PIAT-R to 60% on the DAB and 68% on the WJ-PEB. In all other batteries, except one, fewer than 20% of all KR-20 reliability estimates reported were below.80. The WRAT-R did not report any KR-20 reliability estimates. With the exception of the CAT-E and ITBS, 90% or more of the the reliability estimates reported for group tests were above .80.

Insert Table 1 about here

Item Classifications

The percent of items in each battery that were judged to sample each of the nine categories are shown in Table 2. Test battery names are listed in the left column: other columns correspond to each of the skill categories (listening comprehension, reading decoding, reading comprehension, study skills, language mechanics, language expression, spelling, math computation, and math applications). The values in the body of the table indicate, for each test,

the percent of all items, across all levels of the battery that were categorized in the particular skill area. For example, on the CAT E, 1 percent of all items were categorized as measures of listening comprehension, 7 percent of all items were categorized as measures of reading decoding, 28 percent of all items were categorized as measures of reading comprehension, and so on.

Insert Table 2 about here

The proportion of items classified in each of the skill categories tended to be similar from test to test, particularly on the group administered batteries. By far, the largest proportion of items on group tests were categorized as measures of reading comprehension (generally about 25%). Generally, about 15% of all items on group tests were categorized in each of the Math Application, Math Computation, and Language Mechanics categories. Representation in all other categories was around 5-10%. The distribution of items across categories on individual tests was less distinct with the exception that Reading Decoding accounted for about 1/3 of all items on 3 out of 4 of the batteries. None of the items on individual batteries were classified as measures of study skills or language expression and only 2 out of 4 of the individually administered batteries included listening comprehension items.

The extent to which subtests were judged to measure the skills they purported to measure is shown in Table 3. Battery names and subtest clusters are displayed in the left column of the table. Group tests are shown in the top portion of the table. Skill categories are shown in the 9 columns in the middle of the table, with construct groupings separated by bold vertical lines. The values in the body of the table indicate the percent of items from each subtest cluster categorized in each skill category. For example, on the CAT-E, 2% of items in the reading subtests cluster from all levels of the battery were classified as measures of Listening Comprehension. Similarly, 17% of CAT-E items in the reading subtests cluster were classified as tests of Reading Decoding, 69% were measures of Reading Comprehension and 12% were classified as measures of Study Skills.

Insert Table 3 about here

The right column in Table 3 indicates the percent of items from each subtest cluster placed in one of the skill categories that comprise an appropriate construct grouping. On the CAT-E, for example, 98% of items from the reading subtests cluster were categorized as measures of reading in either the Reading Decoding, Reading Comprehension, or Study Skills categories. A number of test batteries did not include tests intended to assess listening comprehension and these are indicated with NA in the corresponding cells in the table.

Generally, on group tests, items were classified as measures of the construct they purported to test. With one exception, all values in the right column for group tests were above 90%. On the MAT-6, only 80 items from the language subtests cluster were classified as measures of the language construct (i.e., placed in either the Language Mechanics, Language Expression or Spelling Categories). The remainder of language subtests cluster items were judged to be measures of Listening Comprehension (11%) or Study Skills (9%).

Individually administered tests showed a similar but less distinct pattern of congruence between the construct tests purported to measure and the construct they were actually judged to measure. On the PIAT-R and the WJ-PEB, about 20% of language subtest cluster items were categorized as measures of Reading Decoding, however, all other items on individually administered tests were classified in the appropriate construct category.

Some general trends in the distribution of items in each subtest cluster across skill categories can be observed in Table 2, particularly in group batteries. Approximately 65% of all reading items on the group tests were judged to measure reading comprehension. About 40% of all math items test computation skills and about 45% of all language items measure language mechanics. On individually administered tests, the pattern of distribution of items is less distinct with considerable variability across batteries on the dimension of distribution of items.

Patterns of Intercorrelations

The results of the convergent-discriminant validity analyses are shown in Table 3. Test names are listed in the left column of the table with ranges and medians for related and unrelated subtests in the body. The percent of overlap of related and unrelated subtest is shown in column on the far right. The TAP and PIAT included only one subtest each in language and mathematics, so no intercorrelations among related subtest were available. Also, the 3-R's reported only one subtest for each skill domain and no intercorrelations were reported for the WRAT-R, therefore, these tests are not shown in this table.

Considerable overlap was observed on all batteries, of the range of intercorrelations of related subtests with that of unrelated subtests. Correlations among related subtests ranged from .38 (DAB: Language) to .98 (W-JPEB Achievement: Language) while correlations among unrelated subtests ranged from .14 (Circus: Reading) to .85 (ITBS: Mathematics) For example, on the ITBS, mathematics related subtest intercorrelations ranged from .57 to .80 while unrelated subtest intercorrelations ranged from .36 to .85.

Insert Table 4 about here

Intercorrelations among unrelated subtests were generally smaller than intercorrelations among related subtests. Unrelated medians ranged from 18% smaller (W-JPEB Achievement: Language) to 20% larger (DAB: Math) than related medians. Most differences were smaller than 15% (n=22), over one third of the differences were smaller than 10% (n=12) and on 9 batteries, a difference of 5% or less was observed. In only one battery, (DAB: mathematics) were median unrelated coefficients more highly correlated than related coefficients. Ironically, although only one related coefficient was reported for DAB: mathematics, this figure was lower than the median of unrelated subtests, resulting in a positive difference value.

The statistical significance of differences was estimated by examining the 95% confidence interval represented by notches on the Tukey box plots. Overlapping notches on adjacent distributions implied no significant difference. The results of this procedure also are shown on Table 4.

Discussion

Published achievement tests have been characterized here as measures of multiple constructs (such as "reading", "mathematics", or "language") that employ separate subtests to sample discrete behaviors in each construct. The structure and scoring of achievement batteries implies that they operate at three different levels; subtest, total test, and battery, with different interpretations associated with scores obtained from each level.

Validity at the subtest level would be supported by evidence that individual subtests measure distinctly different behaviors. Examination of the intercorrelations of related and unrelated subtests revealed that publishers have not provided sufficient evidence to support the premise that distinctly different constructs are represented by individual subtests. Ranges of intercorrelations of related subtests frequently overlapped with those of unrelated subtests indicating that subtests purporting to measure distinctly different constructs were more highly intercorrelated than subtests intended to measure facets of a single construct. This observation was consistent across all batteries, in all three constructs and often, the overlap was substantial or complete. For example on the SRA, the range of intercorrelations among related language arts subtests was .61 to .80 while the range of intercorrelations of language arts subtests with subtests unrelated to language arts was .36 to .84. On the SRA, as well as numerous other batteries, subtests intended to measure facets of the construct language arts could as likely be measuring facets of the construct "mathematics".

Although no standard exists with which to evaluate the magnitude of differences between related and unrelated subtest intercorrelations, the meager differences obtained here do not offer compelling support for the premise that distinctly different constructs are represented by subtests. Differences of under 5% observed in almost all batteries provide further indication that that many of the intercorrelations for related and unrelated subtests are of approximately the same

magnitude. In sum, little evidence was found to support convergent or discriminant validity at the subtest level.

Examination of reliability data suggested that at some levels on some batteries, subtests are unreliable. Interestingly, the percentage of subtests with low reliability doesn't seem to be a function of how many items are presented throughout the different levels of the test. For example the ITBS has a large number of items across all levels of the test (n=3726) (Nolet & Tindal, in preparation) and 19% of all reliability coefficients reported for the ITBS were below .80 while the TAP had fewer than 1000 items with no subtests below .80 reliability. However, the TAP has only 4 subtests while the ITBS has 15 related to reading, language, or math constructs across all levels of the battery. Clearly, reliability is a function of the number of items per subtest, not the total number of items included. The WJ-PEB achievement test had the worst combination of total items (n=265) and number of subtests (n=7). The net effect is a high percentage of subtests (68%) below an acceptable level of reliability. The DAB suffers from the same ineffective combination of too many subtests and not enough items and reported 60% of all reliability coefficients below .80. (Remarkably, both of these tests are individually administered and presumably intended for use as diagnostic or placement tools, requiring a relatively high standard of technical adequacy).

Data presented in the current study suggest that test publishers are illadvised to demarcate many subskills in the categories of reading, language and math. The major categories of reading, language and math may represent the most efficient level of refinement. Most subtests were categorized appropriately into the construct category they were intended to measure. In other words, it didn't matter how many different subtests were included under the construct, they all seemed to be measures of the construct of interest. However, differences among subtests within each construct seem to be related more to the name assigned to the subtest by the publisher than to the extent to which subtests actually sample discrete aspects of the construct.

Comparison across test batteries indicates that there is a differential emphasis in what constitutes a given construct. For example on the WJ-PEB, 75% items in reading are reading decoding with 25% comprehension while in the CAT, 69% of reading items are comprehension

and only 17% are decoding. Such differences might be a function of the theoretical perspective of administration formats (the former is individual and the latter is group administered). However comparison of more comparable tests indicated similar differences. The ITBS compared to the MAT6 reveals distinctly different emphases within an academic domain. On the ITBS, over 1/3 of the reading items were classified as measures of study skills while no reading items were categorized as study skills on the MAT-6. Both of these tests are group administered, multi-skill, multi level. Comparable differences among tests also were observed in language categories of expression mechanics and spelling. While one test emphasizes mechanics, (i.e., ITBS and DAB), others emphasized spelling (i.e., PIAT and Stanford).

In contrast to reading and language, math was more uniformly distributed across the categories of computation and application. The only exceptions were the TAP and the PIAT. However, again these differences may be due to administration or age range considerations (the latter is an individually administered test and the former is intended for secondary level applications).

Distribution of items across skill categories indicated that achievement batteries are primarily tests of reading comprehension, math problem solving, and language mechanics. Consistently, the greatest proportion of all items on group administered batteries were classified in the reading comprehension skill category with math application and language mechanics tied for a distant second.

Both reliability and validity data indicate that with respect to the number of facets into which constructs are divided, fewer is better. Clearly, the number of subtests included in most batteries is not supported. At best, the analyses described in this study supports demarcation of three major areas: reading, math and language arts. Such a structure would generate higher reliability, and could support inferences about learning in broad domains rather than in specific skill domains that may not be reliable or valid. Group achievement batteries then, may best be characterized as broad-band indices of generalized achievement that support few inferences about learning in any facet of a given construct.

The notion of "achievement in general" seems to missing from current conceptualizations of learning and testing. For example in special education, students are thought to have skill specific disabilities and such practices as profile analysis are predicated on skill specific differences. However, when there is shared variance between tests, even two highly reliable tests, reliability of difference scores actually decrease. (Thorndike & Hagan, 1977). Constructs such as reading, math or language expression may involve multiple facets but the tests examined in the current study seem not to be sufficiently successful in sampling these skills to support inferences about skill specific achievement or apparent differences among scores.

Finally, the logic of construct validation assumes adequate domain sampling and avoids the mono-operation bias (Cook & Campbell, 1975; Messick, 1989) which these tests clearly have. The format employed in most published achievement tests forces construct under-representation All but two of the batteries examined relied exclusively on multiple choice selection type responses and the two batteries that did include production responses (DAB and WJPEB) did so rarely. Each of the constructs, reading, language arts and mathematics, includes dimensions that require active production of behaviors. For example, one of the most important outcomes expected of language arts instruction is facility in written expression. Any test of language arts that fails to include a writing sample cannot possibly claim to include all important dimensions of the construct. Similar arguments can be made for the importance of oral reading fluency in the construct reading and problem solving with algorithms in math. Clearly, the tests analyzed in this study fail to represent the wide range of classroom relevant behaviors that are components of each construct.

Issues of convergent and discriminant validity are extremely salient here. For example, a "Total Language" score is based on 2 equally important premises. The first is that specific language arts subtests adequately represent facets of the construct "language arts", and the second is that behaviors other than those related to "language arts" do not influence the scores obtained on language arts subtests. These premises also underlie testing of the constructs of reading and mathematics and therefor bear some discussion.

The former premise relates to the issue of construct under-representation (Cook & Campbell, 1975). This phenomenon occurs when a test is "too narrow and fails to include important dimensions or facets of the construct" (Messick, 1989, p 34). The major implication of construct representativeness is readily apparent. To make valid inferences about a student's performance in a construct of interest on the basis of a test score, the test must fairly represent the construct. If the test is under-representative, inferences can only pertain to the explicit facets of the construct that were actually tested.

The issue of test format relates to the second premise underlying "Total test" scores, i.e., scores are not influenced by behaviors other than those related to the construct of interest. The use of a single response format across subtests forces all behaviors within a battery to look the same. On most batteries, to respond to word attack items, math computation items or language mechanics items, the test taker performs the same task of choosing from among 3 or 4 choices, the answer that best completes the item. This mono-operation bias permits irrelevancies such as test taking behavior and motivation to influence scores. If a single dimension such a response set or previous practice with the test format can influence scores across a variety of subtests or constructs, the validity of inferences based on scores obtained from the test becomes suspect.

What decisions can these tests support? They clearly do not have content validity, weren't meant to and therefore can't be used for planning or evaluating specific instructional strategies. Unfortunately, researchers have focused on this aspect of the tests, consistently finding that these batteries perform poorly a task for which they were not designed.

Most of these tests have more than adequate criterion related validity. In fact their technical manuals indicate that many of these batteries are highly correlated with one another. Here the decision being made is related to which test to use and the primary responsibility for the validity of decisions rests with the test user. A given test should be chosen because it is the best measure for a particular decision. As has been shown here, even though these tests are intercorrelated, they are not comparable and the manner in which any single test measures a construct must be determined on a test by test basis. The real issue is construct validity, i. e. what construct do you want to measure and how do you define it?

The published achievement test batteries examined here seem to have convergent validity but no discriminant validity (Tindal & Nolet, 1990) and mono-operation bias, severely limiting the kinds of inferences that can be made. These tests can serve as a moderately useful and very heavy anchor that doesn't move. At maximum they can be marginally useful for documenting overall program level decisions that provide comparability across widely disparate programs on a national level. With respect to individual student decisions, these tests could provide a standard against which to compare students for the purpose of making gross "low stakes" screening decisions. However, these tests can't provide information to support inferences about the extent to which a particular curriculum works in a particular grade; the effectiveness of a particular teacher, or the outcome of a particular experimental intervention and they can't be ethically used for such purposes.

References

Campbell, D. T, & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 2, 81-105.

Cook, T.D. & Campbell, D. T. (1979). Quasi-Experimental Design and Analaysis Issues for Field Settings. Boston, MA: Houghton Mifflin

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 581-302.

Ebel, R. L. (1978). The case for minimum competency testing. Phi Delta Kappan, 59, 546-549.

Farr, R., & Carey, R. F. (1986). Reading: What can be measured? (2nd Edition). Newark, DE: International Reading Association, Inc.

Freeman, D. J., Kuhs, T. M., Porter, A. C., Floden, R. E., Schmidt, W. H., & Schwille, J. R. (1983). Do textbooks and tests define a national curriculum in elementary school mathematics? Elementary School Journal, 83, 501-513.

Good, R. H., & Salvia, J. (1988). Curriculum bias in published, norm-referenced reading tests: Demonstrable effects. School Psychology Review, 17, 51-60.

Hall, B.W. (1985). Survey of the technical characteristics of published educational achievement tests. Educational Measurement: Issues and Practices, 4 6-14

Jenkins, J., & Pany, D. (1978). Standardized achievement tests: How useful for special education? Exceptional Children, 44, 448-453.

Knifong, J. D. (1980). Computational requirements of standardized word problem tests. Journal for Research in Mathematics Education, 11(1), 3-9.

Lazar-Morrison, C. Polin, L., Moy, R., & Burry, J. (1980). A review of the literature on test use. CSE Report No. 144, Center for the Study of Evaluation. Graduate School of Education. University of California, Los Angeles

Leinhardt, G., & Seewald, A. (1981). Overlap: What's tested, what's taught? Journal of Educational Measurement, 18, 85-96.

Messick, S.. (1981). Evidence and ethics in evaluation of tests. Educational Researcher, November, 9-19

Messick, S.. (1989). Validity. in R. Linn (Ed.) Educational Measurement, third edition. New York: American Council on Education and MacMillan.

Nolet, V. & Tindal, G. (1990). Construct validity in published achievement tests. The Oregon Conference Monograph, 1990 Eugene, OR: College of Education.

Nolet, V. & Tindal, G. (In preparation). Construct validity in published achievement tests: Final research report. Eugene, OR: College of Education.

Petrosko, J. M. (1978). The quality of standardized high school mathematics tests. Journal for Research in Mathematics Education, 9(2), 137-148.

Prescott, G. A., Balow, I. H., Hogan, T. P. & Farr, R. C. (1988). Technical Manual for the MAT6 Survey Battery Orlando, FL: Harcourt Brace, Jovanovich.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening? Phi Delta Kappan. 62, 631-634

Salvia, J., & Ysseldyke, J. (1988). Assessment in special and remedial education. Boston: Houghton-Mifflin.

Shapiro, E. S., & Derr, T. F. (1988). An examination of overlap between reading curricula and standardized achievement tests. The Journal of Special Education, 21(2), 59-67.6.

Tindal, G. & Marston, D. (1990). Classroom Based Assessment Columbus, OH: Merrill Publishing Co.

Tindal, G. & Nolet, V. (1990) The construct validity of curriculum-based measures of achievement: A multitrait-multimethod analysis. Paper presented at the American Educational Research Association, Boston, MA.

Thorndike, R. L. & Hagen, E. (1977). Measurement and Evaluation in Psychology and Education. New York: John Wiley & Sons.

Tukey, J.W. (1977) Exploratory Data Analysis. Reading, MA: Addison-Wesley.

Appendix A: Tests Reviewed

Test Name	Year	Publisher
Group Administered Tests		
California Achievement Tests, Form E (CAT)	1985	CTB/McGraw-Hill
Comprehensive Test of Basic Skills, Form U (CTBS)	1982	CTB/McGraw-Hill
Circus/STEPIII	1979	Addison-Wesley
Iowa Test of Basic Skills (ITBS)	1986	Riverside Publ. Co.
Test of Achievement & Proficiency (TAP)	1986	Riverside Publ. Co.
Metropolitan Achievement Tests , Form 6 (MAT6)	1988	The Psychological Corporation.
SRA Survey of Basic Skills , Form H (SRA)	1985	Scientific Research Associates.
Stanford Achievement Test Series	1985	The Psychological Corporation.
Individually Administered Tests		
Diagnostic Achievement Battery (DAB)	1984	Pro-Ed
Woodcock-Johnson Psycho-Ed Battery (WJPEB)	1977	Teaching Resources
Peabody Individual Acheivement Test, Revised (PIAT)	1988	American Guidance Service
Wide Range Achievment Test , Revised (WRAT-R)	1984	Jastak Assessment Systems

Test Battery	Total Reported	%< .80
Group Tests		
CAT E	207	14
CTBS-U	197	10
Circus/Step	301	8
ITBS	103	19
Mat 6	363	8
SRA	241	5
Stanford	263	1
TAP	16	0
3-R's	36	6
Individual Tests		
DAB	81	60
PIAT R	27	0
WRAT R	0	
W-JPEB	56	68

Table 1 . Total number of reliability coefficients (KR-20) reported for each battery and percent below .80.

Test Battery	LC	RD	RC	SS	LM	LE	SP	MC	MA
Group Tests									
CAT E	1	7	28	5	15	8	9	13	15
CTBS-U	2	8	27	5	15	8	9	13	14
Circus/Step	11	6	23	8	9	5	3	14	20
ITBS	3	5	22	17	24	3	8	9	11
MAT 6	5	8	31	2	11	1	8	13	21
SRA	2	6	28	9	12	2	9	13	19
Stanford	10	19	26	3	5	3	11	9	13
TAP	0	0	27	27	12	12	2	3	16
3-R's	0	3	31	7	19	3	6	14	17
Individual Tests									
DAB	20	8	17	0	30	0	6	10	8
PIAT R	0	30	24	0	2	0	20	5	19
WRAT	3	33	0	0	0	0	30	33	0
W-JPEB	0	36	10	0	11	0	9	16	18

Table 2 Percent of items classified in each skill category for each test battery

Codes: LC: listening comprehension, RD: reading decoding, RC: reading comprehension, SS: study skills, LM: language mechanics, LE: language expression, SP: spelling, MC: math computation, MA: math applications

	LC	Reading			Language			Math		% App Const
Group Administered Tests	LC	RD	RC	SS	LM	LE	SP	MC	MA	
CAT E										
% list comp subtest items	100									100
% reading subtest items	2	17	69	12						98
% language subtest items					47	25	27			100
% math subtest items								46	54	100
CTBS U										
% list comp subtest items	100									100
% reading subtest items	5	19	65	11						95
% language subtest items					46	24	29			100
% math subtest items								47	53	100
Circus/Step										
% list comp subtest items	100									100
% reading subtest items		18	64	23						100
% language subtest items		10			48	26	16			90
% math subtest items								40	60	100
ITBS										
% list comp subtest items	100									100
% reading subtest items		12	50	38						100
% language subtest items					70	8	22			100
% math subtest items								45	55	100
MAT 6										
% list comp subtest items	na									na
% reading subtest items	6	19	76							94
% language subtest items	11			9	43	4	33			80
% math subtest items								39	61	100
SRA										
% list comp subtest items	100									100
% reading subtest items		14	65	20						100
% language subtest items					53	7	40			100
% math subtest items								41	59	100
Stanford										
% list comp subtest items	91		9							91
% reading subtest items		43	60	6						100
% language subtest items					23	15	50			100
% math subtest items								40	60	100

Table 3 Percent of items in each subtest cluster classified in skill categories

Codes: LC: listening comprehension, RD: reading decoding, RC: reading comprehension, SS: study skills, LM: language mechanics, LE: language expression, SP: spelling, MC: math computation, MA: math applications

Table 3 (continued)

	LC	Reading			Language			Math		% App Const
Group Administered Tests	LC	RD	RC	SS	LM	LE	SP	MC	MA	
TAP										
% list comp subtest items	na									na
% reading subtest items		0	50	50						100
% language subtest items					46	44	9			100
% math subtest items								17	83	100
3-R's										
% list comp subtest items	na									na
% reading subtest items		8	75	17						100
% language subtest items					68	12	20			100
% math subtest items								45	55	100
Individually Administered Tests										
DAB										
% list comp subtest items	100									100
% reading subtest items		33	67							100
% language subtest items					84		16			100
% math subtest items								55	45	100
PIAT										
% list comp subtest items	na									na
% reading subtest items		50	50							100
% language subtest items		21			7	1	71			79
% math subtest items								19	81	100
W-J PEB										
% list comp subtest items	na									na
% reading subtest items		75	25	0						100
% language subtest items		20			44	0	36			80
% math subtest items								46	54	100
WRAT										
% list comp subtest items	na									na
% reading subtest items		100								100
% language subtest items							100			100
% math subtest items								100		100

Table 3 Percent of items in each subtest cluster classified in skill categories

Codes: LC: listening comprehension, RD: reading decoding, RC: reading comprehension, SS: study skills, LM: language mechanics, LE: language expression, SP: spelling, MC: math computation, MA: math applications

	Related Range	Related Median	Unrelated Range	Unrelated Median	Difference Significant?	% Overlap
CAT						
Reading	.62 -.84	0.72	.45 -.82	0.68	yes	98
Language	.64 -.80	0.69	.46 -.82	0.68	no	100
Mathematics	.62 -.83	0.72	.45 -.82	0.62	yes	91
Circus/STEP						
Reading	.74 -.85	0.79	.40-.83	0.72	yes	94
Language			.14 -.83	0.71	na	na
Mathematics	.63 -.82	0.72	.40 -.78	0.68	yes	98
CTBS						
Reading	.64 - .82	0.77	.44 -.80	0.67	yes	99
Language	.54 -.78	0.67	.45 -.76	0.64	yes	99
Mathematics	.59 -.77	0.70	.43 -.77	0.61	yes	100
ITBS						
Reading	.47 -.83	0.71	.42 -.76	0.63	yes	88
Language	.45 -.76	0.68	.37 -.80	0.60	yes	100
Mathematics	.57 -.80	0.65	.36 -.85	0.60	yes	100
MAT 6						
Reading	.70 -.84	0.77	.42 -.79	0.63	yes	70
Language	.59 -.68	0.63	.44 -.80	0.65	no	100
Mathematics	.54 -.79	0.66	.42 -.79	0.58	yes	100
SRA						
Reading	.44 -.84	0.73	.34 -.84	0.65	yes	100
Language	.61 -.80	0.70	.36 -.83	0.68	yes	100
Mathematics	.54 -.83	0.74	.33 -.79	0.63	yes	68
Stanford						
Reading	.57 -.85	0.71	.42 -.83	0.64	yes	91
Language	.72 -.76	0.74	.43 -.83	0.66	yes	100
Mathematics	.66 -.81	0.72	.42 -.84	0.63	yes	100
TAP						
Reading	.79-.82	0.80	.68 -.82	0.75	yes	100
Language			.70 -.78	0.73	na	na
Mathematics			.68 -.80	0.73	na	na
DAB						
Reading	.81 -.85	0.81	.36 -.83	0.69	yes	66
Language	.38 -.70	0.61	.33 -.80	0.60	no	100
Mathematics	.55 -.55	0.55	.27 -.76	0.66	na	100
PIAT-R						
Reading	.42 -.99	0.73	.43 -.90	0.70	no	93
Language			.30 -.90	0.66	na	na
Mathematics			.35 -.71	0.68	na	na
W-J PEB: Achmt						
Reading	.52 -.80	0.67	.37 -.82	0.64	no	100
Language	.59 -.98	0.77	.41 -.80	0.63	yes	78
Mathematics	.53 -.74	0.66	.37 -.67	0.54	yes	80

Table 4. Summary data for 11 test batteries